

Spreading of intolerance under economic stress: results from a model with reputation.

Luis A. Martinez-Vaquero^{1,2} and José A. Cuesta^{1,3}

¹*Grupo Interdisciplinar de Sistemas Complejos (GISC), Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Madrid, Spain*

²*Current address: Artificial Intelligence Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium*

³*Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, 50009 Zaragoza, Spain*

When a population is engaged in successive prisoner's dilemmas, indirect reciprocity through reputation fosters cooperation through the emergence of moral and action rules. A simplified model has recently been proposed where individuals choose between helping or not others, and are judged good or bad for it by the rest of the population. The reputation so acquired will condition future actions. In this model, eight strategies (referred to as 'leading eight') enforce a high level of cooperation, generate high payoffs and are therefore resistant to invasions by other strategies. Here we show that, by assigning each individual one out of two labels that peers can distinguish (e.g., political ideas, religion, skin colour...) and allowing moral and action rules to depend on the label, intolerant behaviours can emerge within minorities under sufficient economic stress. We analyse the sets of conditions where this can happen and also discuss the circumstances under which tolerance can be restored. Our results agree with empirical observations that correlate intolerance and economic stress, and predict a correlation between the degree of tolerance of a population and its composition and ethical stance.

PACS numbers: 87.23.Ge, 02.50.Le, 89.75.Fb, 87.10.Mn

I. INTRODUCTION

Different kinds of discriminatory behaviour based on membership to different groups—identified, for example, by visible tags [1]—has been reported both in human [2–5] and in animal societies [6–10]. They can be broadly classified into two main types: in-group favoritism and out-group hostility [11–13]. Recent studies of in-group and out-group cooperation show that in-group favouritism can emerge in the framework of indirect reciprocity [14–16].

The concept of indirect reciprocity [17, 18] has been introduced to explain the emergence of cooperation in society, where many interactions between the same individuals have low chances to be repeated. Contrary to direct reciprocity [19], indirect reciprocity implies that individuals receive the consequences of their actions, not directly from their opponents, but indirectly through society. Indirect reciprocity—and its related reputation concept—has proven to be an important mechanism for the emergence and sustainment of cooperation in small-scale human [20–24] and non-human [25] societies. It also plays an important role in communication networks [26, 27]. In these reputation-based models, individuals have an opinion about every interaction they witness and assign a reputation to the individuals involved accordingly [21, 28–31]. The nature of their future interactions with those individuals will be determined by the reputation they have assigned to them.

In a stylised model of indirect reciprocity Ohtsuki and Iwasa [29] and Brandt and Sigmund [30] classified the strategies involved in these games attending to action and assessment modules. The action module prescribes what to do—whether cooperate or not—given the reputations of the individuals involved. The assessment module assigns reputation to the interacting individuals according to a moral code. To do it three elements can be judged: the nature of the interaction, the reputation of the recipient, and the reputation of the donor. Strategies are thus classified as first, second, or third order depending

on whether the first, the first and the second, or all the three elements are taken into account to assign a reputation to the donor. Ohtsuki and Iwasa [29] studied the evolutionarily stability of third order strategies which share the same moral assessment module, and later Martinez-Vaquero and Cuesta [32] extended the study confronting strategies with different moral codes. Among the evolutionary stable strategies (ESS) found, there are eight strategies—the so-called *Leading Eight*—that are considerably more efficient (get very high payoff) and coherent (in terms of consistency between their moral and action modules) than the rest of them. All these strategies foster cooperation through indirect reciprocity.

This simple model of indirect reciprocity can be readily extended to study the emergence of discriminatory behaviours by introducing different groups of individuals identified by external signs (these may include physical or cultural traits). Our goal in this work is to analyse under which conditions an intolerant strategy (understood as out-group hostility) can invade a tolerant population that follows one of the Leading Eight strategies. Whenever intolerance spreads, we are also interested in whether tolerance can be restored by introducing some kind of external incentives. Our hope is that this study sheds some light into the causes for the emergence of intolerance in societies and the mechanisms to alleviate this social burden.

The paper is structured as follows. In section II we introduce the model, which is mathematically implemented in section III. Our results are shown in section IV and discussed in section V.

II. MODEL

Our model is an extension of that used in a previous work [32]—itself a modification of the indirect reciprocity model based on reputation introduced by Brandt and Sigmund [30] and later investigated by Ohtsuki and Iwasa [29, 33].

We consider an infinite, well-mixed population, where every individual is aware of every action performed and produces a moral judgment—which leads to a reputation assignment—on it. Each time step a pair of individuals is randomly drawn—with equal probability—from the population. One of them plays as *donor* and the other one as *recipient*. The donor can decide to pay a cost $c > 0$ to help (C) the recipient or not (D). Any helped recipient gains a payoff $b > c$; otherwise there is no payoff whatsoever.

This *action* is observed by every individual in the population (including themselves). Thus everyone makes a private judgment of the donor for this action. The observer's *moral assessment* will decide the reputation—good (G) or bad (B)—she assigns to the donor. Accordingly, everybody has a private opinion of every other individual—even of herself.

By repeating this process the population eventually reaches a steady state. The average payoff of this repeated game is then computed for every individual. By assuming a very large population we effectively neglect direct reciprocity—the probability that two individuals meet again is very small.

Strategies are defined by two modules: the action rules, which tells the donor how to interact with the recipient, and the moral rules, which prescribe a reputation for the donor of every witnessed action. As in our previous paper [32], we will consider third order strategies.

The action rules determine what the donor must do (either help or refuse to help) given the reputation of both players. Specifically, $a_{\alpha\beta} = 1$ (C) if a strategist with reputation α helps an individual with reputation β and 0 (D) otherwise.

The moral assessments tell the individual if the action just witnessed should be judged good or bad, hence revising the donor's reputation. Specifically, $m_{\alpha\beta}(a) = 1$ (G) if the observer assigns good reputation to a donor she previously judged α , who performs an action a on an recipient she previously judged β , and is 0 (B) otherwise.

Thus each strategy is defined by 12 numbers: 4 for the action module and 8 for the moral module. This amounts to 4096 different strategies in total.

Besides, we will assume that players sometimes make errors when trying to help another individual [29, 34–37]. Thus, with a probability ϵ a donor defects regardless of her action rules and with $1 - \epsilon$ she performs the action she planned to. One possible source of this error—which will be important in interpreting the results of this work—is scarcity of resources. Despite her willingness to help, an individual may fail to do it because she cannot afford the cost c . Errors due to misjudgements are excluded (see [32] for a justification).

Discrimination requires at least two different subpopulations that can be clearly identified by everybody. Thus we divide our population into two classes, A and B. The fraction of A individuals will be denoted y . A *tolerant* player will act disregarding the individuals' class when deciding her action or judgement. An *intolerant* individual acts and judges like a tolerant one with the following exceptions:

1. Individuals of the other class are never helped.
2. Donors of the other class are judged bad under any circumstance.

3. Individuals of her own class who do not help individuals of the other class are always judged good.
4. Individuals of her own class who do help individuals of the other class are always judged bad.

III. IMPLEMENTATION

Our goal is to determine, with the assumptions of the model, under which conditions a small population of intolerant (tolerant) mutants M can invade a tolerant (intolerant) resident population R made of A-type and B-type individuals in fractions y and $1 - y$ respectively. This fraction y will be kept constant and is therefore a parameter of the model. We will only consider scenarios where all individuals of the same class in R show the same kind of behavior—either tolerant or intolerant—toward the other class. Likewise, we will limit ourselves to determine when the resident population can resist the invasion of mutants. Determining the final composition of the population if the invasion succeeds is a more complex problem that we will tackle in this paper.

Without loss of generality, in the following calculations mutants will be assumed to belong to class A. Likewise, we will only need to consider whether A-residents imitate the mutant or not. If a B-resident imitates the A-mutant (its tolerant or intolerant character) this can be treated as an invasion of B-residents by a B-mutant separately—the reason being that mutants are present in so small a fraction that the probability that two mutants interact is negligible.

Accordingly the different invasion scenarios will be denoted $t_M|t_A t_B$, where t_X can be either T (tolerant individuals) or I (intolerant individuals), M denotes A-type mutants, and A and B the two subpopulations. The four possible scenarios are:

- I|TT: intolerant A-mutants try to invade a tolerant population;
- T|IT: tolerant A-mutants try to invade a population made of intolerant A-residents and tolerant B-residents;
- I|TI: intolerant A-mutants try to invade a population made of tolerant A-residents and intolerant B-residents;
- T|II: tolerant A-mutants try to invade an intolerant population.

The condition for A-residents to resist the invasion of A-mutants is $W(A|R) > W(M|R)$, where $W(A|R)$ and $W(M|R)$ are the average payoffs received by A-type residents and by mutants, respectively. They can be obtained as

$$\begin{aligned}
 W(A|R) &= b \Theta_{R,A} - c \Theta_{A,R}, \\
 W(M|R) &= b \Theta_{R,M} - c \Theta_{M,R}, \\
 \Theta_{R,A} &= y \theta_{A,A} + (1 - y) \theta_{B,A}, \\
 \Theta_{A,R} &= y \theta_{A,A} + (1 - y) \theta_{A,B}, \\
 \Theta_{R,M} &= y \theta_{A,M} + (1 - y) \theta_{B,M}, \\
 \Theta_{M,R} &= y \theta_{M,A} + (1 - y) \theta_{M,B},
 \end{aligned} \tag{1}$$

where $\theta_{i,j}$ is the probability that an i -strategist helps a j -strategist. If i -strategists are intolerant and j -strategists

belong to a different class $\theta_{i,j} = 0$; otherwise these probabilities can be obtained as

$$\theta_{i,j} = (1 - \epsilon) \sum_{\alpha\beta} \chi_\alpha(g_i^j) \chi_\beta(g_j^i) a_{\alpha\beta}, \quad (2)$$

where g_i^j is the fraction of i -strategists that are considered good by the j -strategists. We have introduced the auxiliary function

$$\chi_\gamma(x) = \gamma x + (1 - \gamma)(1 - x). \quad (3)$$

Thus if x represent the fraction of good individuals, then $\chi_1(x) = x$ and $\chi_0(x) = 1 - x$ are the fraction of ‘good’ and ‘bad’ individuals, respectively.

In order to calculate the fractions g_i^j we first need to compute the fractions $x_i^{\Lambda_A \Lambda_B \Lambda_M}$ of i -individuals with reputations Λ_A , Λ_B , and Λ_M according to A-residents, B-residents, and mutants, respectively. Here Λ_i can be G (for good reputation), B (for bad reputation), or * (for any reputation, G or B). For instance, x_i^{G*B} represents the fraction of i -type individuals who are considered good by A-residents and bad by mutants, regardless of B-residents’ opinion. Hence $x_i^{G*B} = \sum_{\Lambda_B} x_i^{G\Lambda_B}$. Thus

$$g_i^A = x_i^{G**}, \quad g_i^B = x_i^{*G*}, \quad g_i^M = x_i^{**G}. \quad (4)$$

It turns out that not all $x_i^{\Lambda_A \Lambda_B \Lambda_M}$ are necessary to compute g_i^j (see Appendix for details). In general, all we need is to obtain $x_A^{\Lambda_A * \Lambda_M}$, $x_M^{\Lambda_A * \Lambda_M}$, and for some scenarios also $x_B^{\Lambda_A * \Lambda_M}$ or $x_B^{* \Lambda_B \Lambda_M}$. Since we are considering an infinite population, self-interactions are negligible. Under the assumption that the rate of mutants is very low, donors will only interact with residents and therefore the dynamics of these fractions will be given by

$$\frac{dx_A^{\Lambda_A * \Lambda_M}}{dt} = y T_{A,A}^{\Lambda_A \Lambda_M} + (1 - y) F_{A,B}^{\Lambda_A \Lambda_M} - x_A^{\Lambda_A * \Lambda_M}, \quad (5)$$

$$\frac{dx_B^{\Lambda_A * \Lambda_M}}{dt} = (1 - y) T_{B,B}^{\Lambda_A \Lambda_M} + y F_{B,A}^{\Lambda_A \Lambda_M} - x_B^{\Lambda_A * \Lambda_M}, \quad (6)$$

$$\frac{dx_B^{* \Lambda_B \Lambda_M}}{dt} = (1 - y) T_{B,B}^{\Lambda_B \Lambda_M} + y F_{B,A}^{\Lambda_B \Lambda_M} - x_B^{* \Lambda_B \Lambda_M}, \quad (7)$$

$$\frac{dx_M^{\Lambda_A * \Lambda_M}}{dt} = y T_{M,A}^{\Lambda_A \Lambda_M} + (1 - y) F_{M,B}^{\Lambda_A \Lambda_M} - x_M^{\Lambda_A * \Lambda_M}, \quad (8)$$

where the interactions with recipients of the same and of the opposite class have been split into $T_{i,j}^{\Lambda_i \Lambda_m}$ and $F_{i,j}^{\Lambda_i \Lambda_m}$, respectively. Fractions $F_{i,j}^{\Lambda_i \Lambda_m}$ depend on the scenario we are considering whereas $T_{i,j}^{\Lambda_i \Lambda_m}$ are common for all of them and can be obtained as

$$T_{i,A}^{\Lambda_A \Lambda_M} = \sum_{\alpha\beta\gamma\delta} x_i^{\alpha\beta\gamma} x_A^{\delta\gamma\delta} \times R^{\Lambda_A \Lambda_M}(\alpha\beta\gamma, \alpha\delta\gamma|\alpha_i\beta_i) \quad (9)$$

$$T_{B,B}^{\Lambda_A \Lambda_M} = \sum_{\alpha\beta\gamma\delta} x_B^{\alpha\beta\gamma} x_B^{\delta\gamma\delta} \times R^{\Lambda_A \Lambda_M}(\alpha\beta\gamma, \alpha\delta\gamma|\alpha_B\beta_B) \quad (10)$$

$$T_{B,B}^{\Lambda_B \Lambda_M} = \sum_{\alpha\beta\gamma\delta} x_B^{\alpha\beta\gamma} x_B^{\delta\gamma\delta} \times R^{\Lambda_B \Lambda_M}(\alpha\beta\gamma, \alpha\delta\gamma|\alpha_B\beta_B) \quad (11)$$

where $R^{\Lambda_l \Lambda_m}(\alpha_l \beta_l, \alpha_m \beta_m | \alpha_i \beta_i)$ is the probability that an i -type donor with reputation (α_l, α_m) acting on a recipient with reputation (β_l, β_m) , is assigned a reputation (Λ_l, Λ_m) . Reputations are as given by l -individuals and m -individuals respectively. This probability is obtained as

$$\begin{aligned} R^{\Lambda_l \Lambda_m}(\alpha_l \beta_l, \alpha_m \beta_m | \alpha_i \beta_i) &= \\ &= (1 - \epsilon) \delta(\Lambda_l, m_{\alpha_l \beta_l}(a_{\alpha_i \beta_i})) \delta(\Lambda_m, m_{\alpha_m \beta_m}(a_{\alpha_i \beta_i})) \\ &+ \epsilon \delta(\Lambda_l, m_{\alpha_l \beta_l}(D)) \delta(\Lambda_m, m_{\alpha_m \beta_m}(D)). \end{aligned} \quad (12)$$

In general, Eqs. (5)–(8) need to be solved numerically. Nevertheless, for the different scenarios that we will consider, some simplifications can be made. The details are deserved to the Appendix.

IV. RESULTS

We focus our analysis on the *leading eight* strategies introduced by Ohtsuki and Iwasa [29]. In a previous paper [32] we proved that, for reasonable values of benefit-to-cost and error rates, these eight strategies are evolutionarily stable against the invasion of any other strategy. On the other hand, Ohtsuki and Iwasa [29] classified the leading eight in three groups (Table IV), according to their stability against invasions by fully defective action rules (AllD). Neglecting errors in the moral assessment, Groups I and II can be invaded by AllD provided

$$\frac{b}{c} < 1 + \epsilon + \epsilon^2 + \dots = \frac{1}{1 - \epsilon}, \quad (13)$$

in other words, if the error in the action $\epsilon > (b - c)/b$. Invading Group III requires $b < c$ though.

Therefore, for a given $b > c$, if ϵ is sufficiently large, any Group I or Group II strategy can be invaded by defectors, whereas Group III strategies always resist invasions. As we will show below, these are the same conditions under which intolerance can spread, but intolerant strategies obtain a higher payoff than AllD because they are indistinguishable from the strategy used by residents of their same group. Accordingly, when two distinguishable groups coexist in a population, intolerant strategies are preferred over AllD.

In what follows we will consider the different invasion scenarios for which there exists a parameter region where invasion occurs. These regions are determined numerically, and represented in Figs. 1–8 as colored areas in a ϵ – y plane, where y represents the fraction of A-residents in the community. For the sake of clarity of the representations, we will sometimes assume that the invader is of type A, and sometimes of type B. Consistent with the previous section, we will denote $t_M | t_{ATB} - t_X = T, I$ being the tolerant/intolerant character of type X individuals—whenever the mutant is of type A, and $t_{ATB} | t_M$ whenever the mutant is of type B. (Notice, however, that y will always represent the fraction of A-residents, which will coincide with the type of mutants in the former case, but not in the latter.) Changing the type of mutant will make analyzing successive invasions $I|TT$ and $IT|I$ —which would turn a fully tolerant population into a fully intolerant one—much easier.

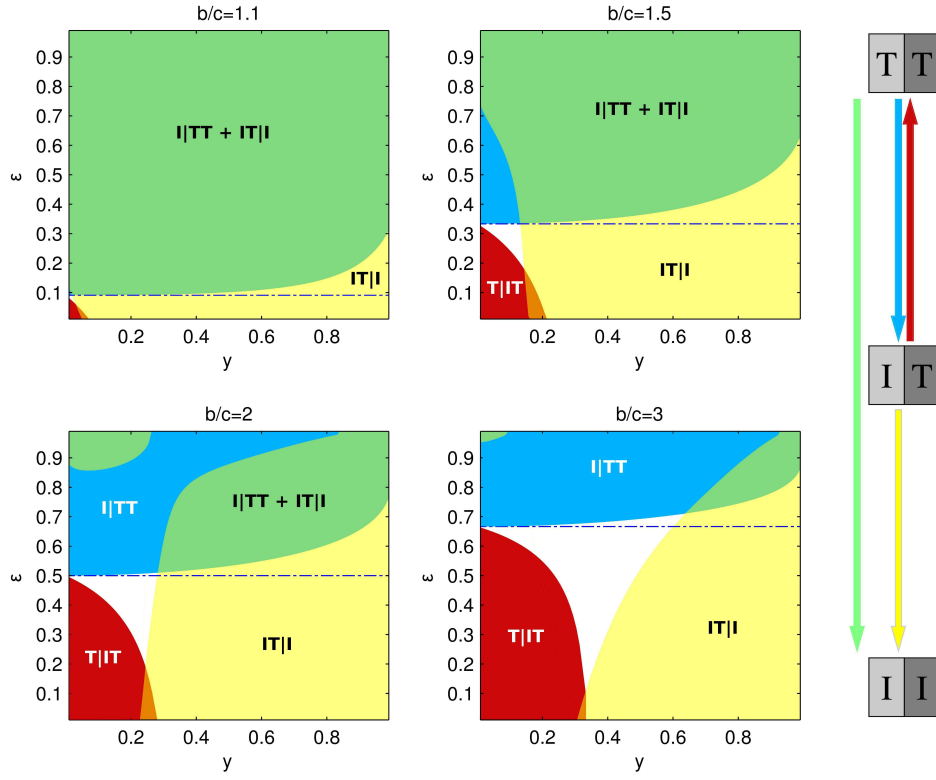


FIG. 1. (Color online.) Values for ϵ and y indicating the different regions where invasions $I|TT$, $T|IT$ and $IT|I$ are successful for strategy Ia. Above the horizontal line, also AllD can invade a tolerant population.

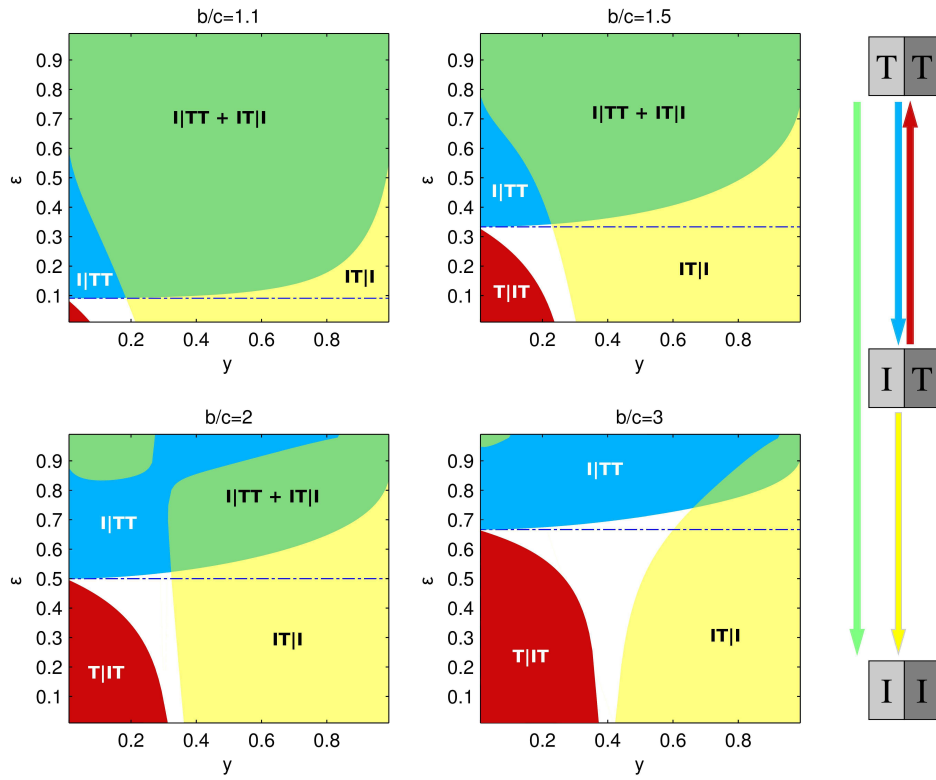


FIG. 2. (Color online.) Same as Fig. 1 but for strategy Ib.

	$m_{GG}(C)$	$m_{GG}(D)$	$m_{GB}(C)$	$m_{GB}(D)$	$m_{BG}(C)$	$m_{BG}(D)$	$m_{BB}(C)$	$m_{BB}(D)$	a_{GG}	a_{GB}	a_{BG}	a_{BB}
Ia	G	B	G	G	G	B	G	B	C	D	C	C
Ib	G	B	B	G	G	B	G	B	C	D	C	C
IIa	G	B	G	G	G	B	G	G	C	D	C	D
IIb	G	B	G	G	G	B	B	G	C	D	C	D
IIc	G	B	B	G	G	B	G	G	C	D	C	D
IId	G	B	B	G	G	B	B	G	C	D	C	D
IIIa	G	B	G	G	G	B	B	B	C	D	C	D
IIIb	G	B	B	G	G	B	B	B	C	D	C	D

TABLE I. Leading eight strategies. The first eight columns describe the moral assessment; the last four, the action rules. Strategies are classified into three groups (I, II, and III; see text).

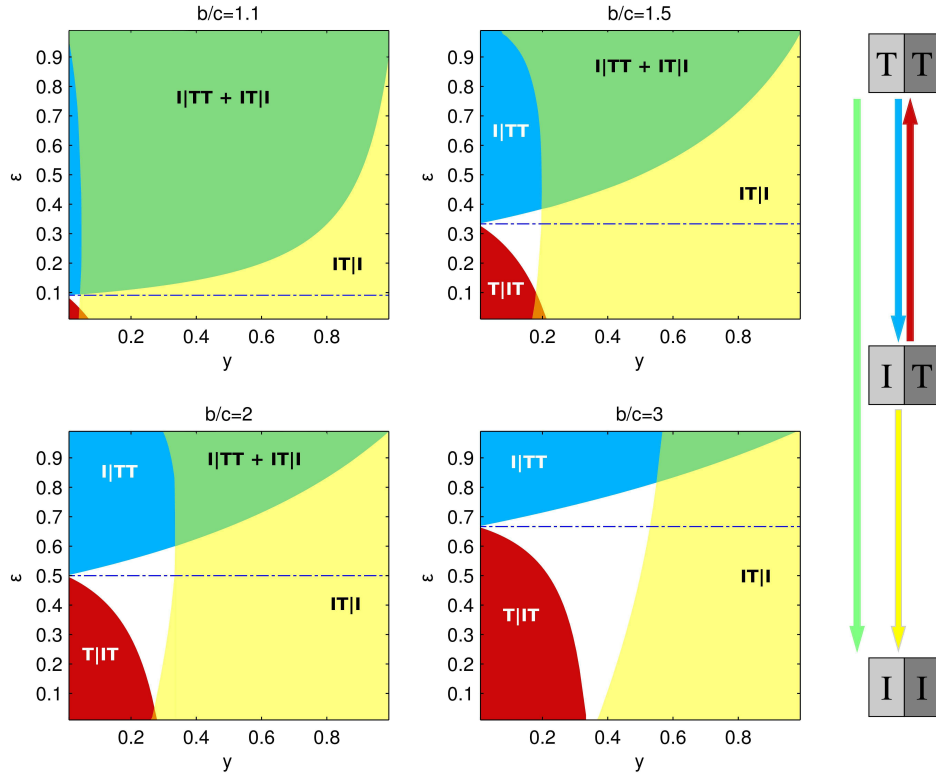


FIG. 3. (Color online.) Same as Fig. 1 but for strategy IIa.

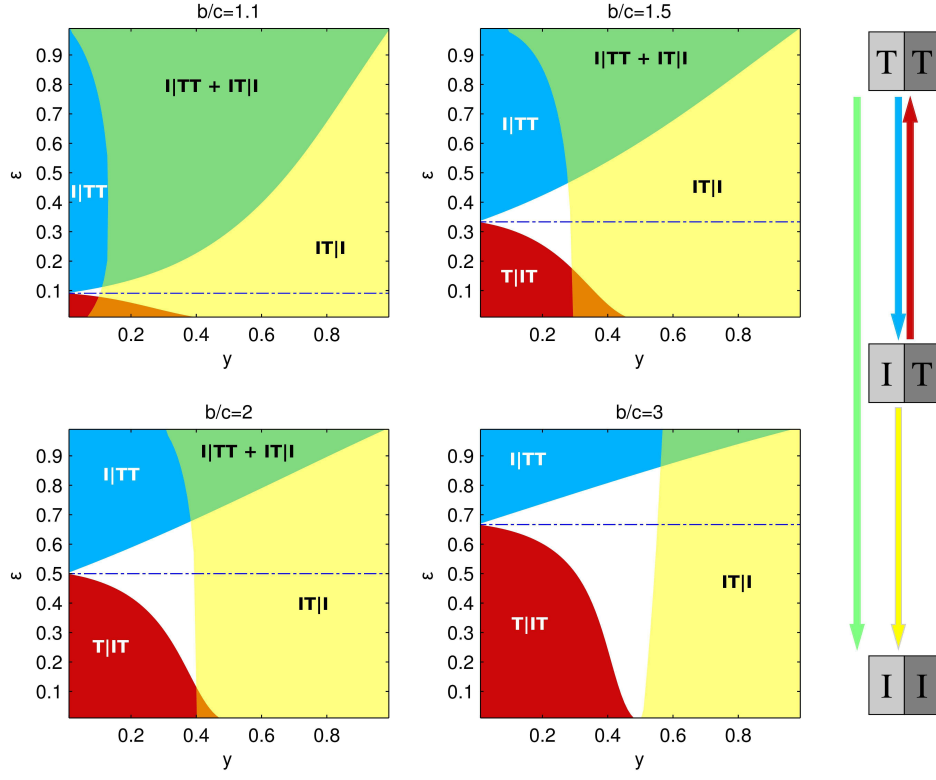


FIG. 4. (Color online.) Same as Fig. 1 but for strategy IIb.

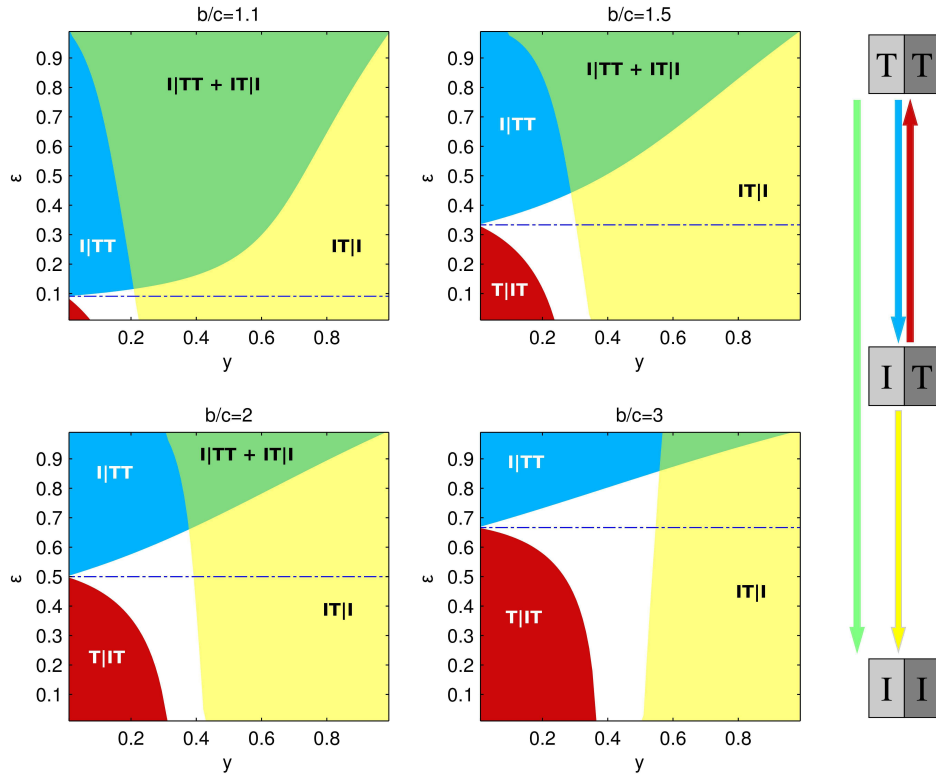


FIG. 5. (Color online.) Same as Fig. 1 but for strategy IIc.

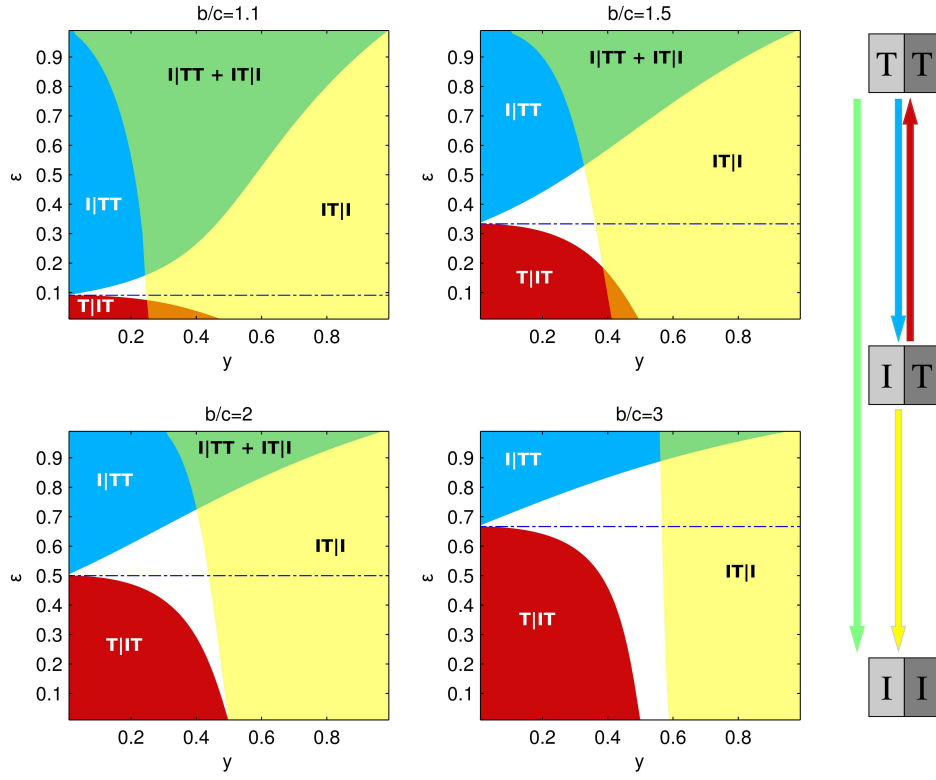


FIG. 6. (Color online.) Same as Fig. 1 but for strategy II d.

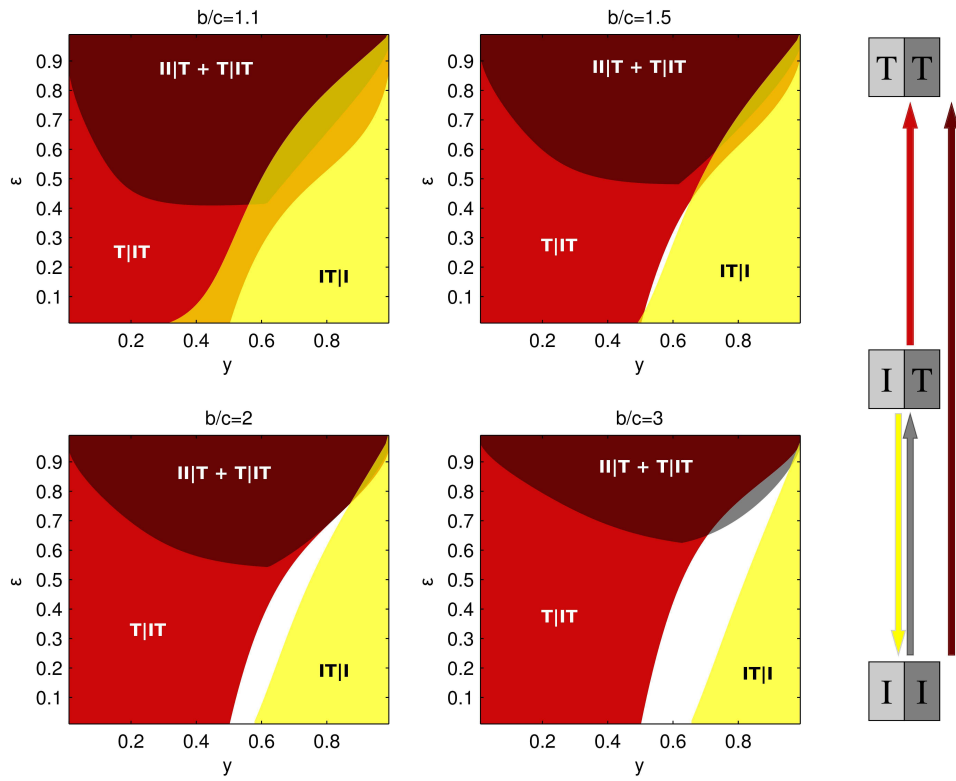


FIG. 7. (Color online.) Same as Fig. 1 but for strategy III a, with the new invasion scenario $II|T$.

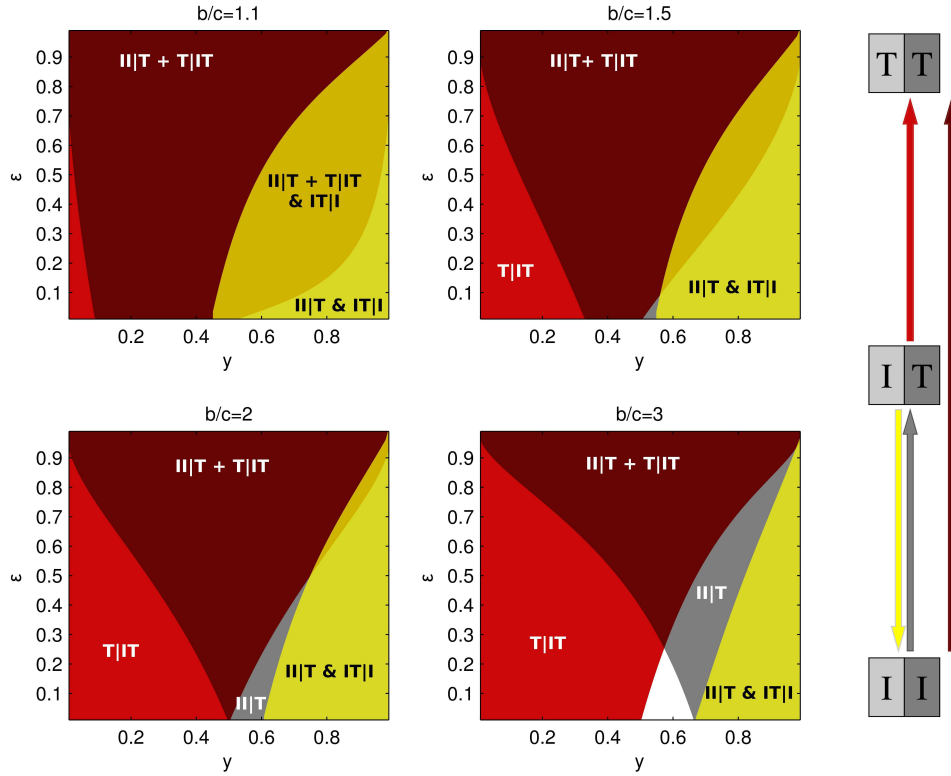


FIG. 8. (Color online.) Same as Fig. 7 but for strategy IIIb.

In all invasion scenarios that we will consider, we will just determine the stability of the resident population against invasion by mutants, but not the final equilibrium (hence we do not determine whether there is a turnover of the invaded population or a final coexistence is reached). Figures 1–6 correspond to Group I and Group II strategies, whereas Figs. 7–8 correspond to Group III strategies.

The different scenarios are as follows: $I|TT$, $T|IT$, $IT|I$, and $T|II$. Overlaps of these regions are colored by a mixed color.

First of all, Figures 1–6 show that tolerant communities that follow Group I and Group II strategies can be invaded by intolerant individuals ($I|TT$) if ϵ is sufficiently high and/or b/c sufficiently low. Minorities are more prone to undergo such a spread of intolerance. Second, if one of the two resident types is intolerant and the other one is tolerant, intolerant residents can be invaded by tolerant mutants ($T|IT$) provided they are a minority (in all cases that we have analysed $y < 0.5$). However, this $T|IT$ region never overlaps with the $I|TT$ one, so after a spread of intolerance the “economic conditions” must change for tolerance to be restored. Third, if in a community with tolerant and intolerant residents the latter are the majority, then this strategy can spread also among the tolerant residents ($IT|I$), thus dividing the community into two separate groups that dislike and never help each other. And finally, a fully intolerant community resists invasion by tolerant individuals for any combination of parameters. Hence, there is no way to restore tolerance in a fully intolerant community by tuning the parameters of the model (but see Section IV A).

If we associate low b/c ratios and high ϵ with strait conditions, the interpretation emerging from these results is that economic stress favors the spread of intolerance within minorities. Once it has invaded a group, tolerance cannot be restored unless the economic conditions improve. And if intolerance has eventually split the community, not even that can restore tolerance again.

Group III strategies are quite different (Figs. 7–8). A tolerant population following any of these strategies always resists the invasions of intolerant mutants regardless of the parameter setting (i.e., invasion $I|TT$ never happens). Furthermore, if one type is tolerant and the other one intolerant in the resident population, tolerant mutants can invade the intolerant residents ($T|IT$). Intolerant mutants can also invade the tolerant residents ($IT|I$), but this occurs in a much narrower region of the diagram compared to what happens for Group I and Group II strategies. As a matter of fact, intolerant mutants only succeed if they try to invade a minority of tolerant residents when the rest of the community is intolerant. Furthermore if we start from a society where both subpopulations are intolerant, tolerant mutants have a chance to invade the minority and spread their strategy ($II|T$)—and perhaps eventually turn the community into tolerant individuals in the brown region, overlap between $II|T$ and $T|IT$ invasion regions. Overall, Group III strategies prove more stable against intolerance.

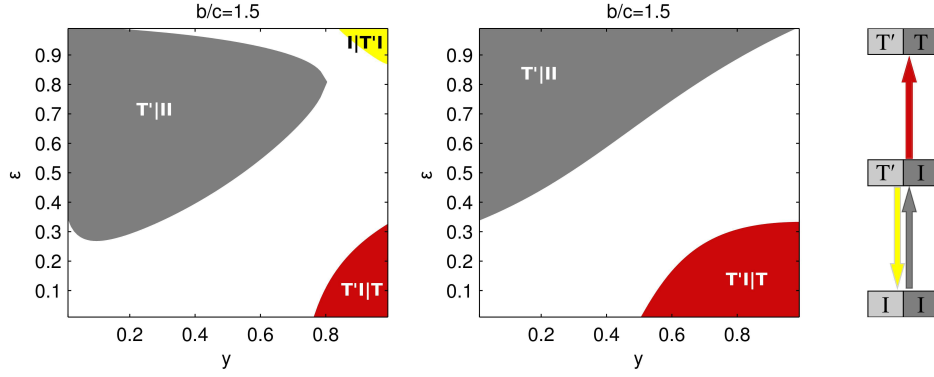


FIG. 9. (Color online.) Conditions of ϵ and y under which the invasions $T'|II$, $I|T'I$ and $T'I|T$ are successful for Ib (left) and IId (right) strategies. Incentivated cost $c' = 0$.

A. Incentives to restore tolerance

As shown in the previous section, if a tolerant community—or one of the two subpopulations within it—following a Group I or a Group II strategy eventually becomes intolerant, there is no way to restore tolerance again by spreading it among the intolerant individuals. The only way to achieve this is by “improving the economic conditions” (i.e., decreasing ϵ and/or increasing b/c). An alternative way to generate the same effect could be introducing an exogenous incentive. We have implemented one such policy through a reduction of the cost of helping others. In other words, we assume the presence of a super-agent who partly subsidises the cost of helping individuals of a subpopulation. This is done by lowering the cost of helping B -strategists. By assuming this cost $c' < c$, Eqs. (1) turn into

$$\begin{aligned} W'(A|R) &= b \Theta_{R,A} - \Theta'_{A,R}, \\ W'(M|R) &= b \Theta_{R,M} - \Theta'_{M,R}, \end{aligned} \quad (14)$$

where

$$\begin{aligned} \Theta'_{A,R} &= c \theta_{A,A} + c' \theta_{A,B}, \\ \Theta'_{M,R} &= c \theta_{M,A} + c' \theta_{M,B}. \end{aligned} \quad (15)$$

The new scenarios where individuals of a class are given incentives to help the other class are $I|T'T$, $T'|IT$, $I|T'I$, and $T'|II$, where primes mark individuals who pay the lower cost c' for helping the opposite class. Equations (14) must be used for these individuals, whereas non-primed individuals follow Eqs. (1), as usual. Intolerant individuals are never marked since they never help the opposite class anyway.

From now on we focus only in Groups I and Group II strategies because tolerant populations following Group III strategies are intolerance-proof. The most effective incentive is reached for $c' = 0$ (i.e., helping the opposite class is free). Since strategies belonging to the same group have a very similar behaviour, we show only one example for each group in Figs. 9-10.

Starting from a fully intolerant population, we represent in Fig. 9 the region where incentives are able to promote tolerance in one of the subpopulation ($T'|II$). Wherever this invasion succeeds, the situation cannot be

reverted, i.e., invasions $I|T'I$ never occur for the same parameter values ($I|T'I$ and regions $T'|II$ do not overlap). But on the other hand, intolerant residents in a $T'I$ population cannot be invaded by tolerant mutants ($T'I|T$) for the same parameter values (again, $T'I|T$ and $T'|II$ regions do not overlap). In other words, if incentives promote tolerance in one subpopulation of fully intolerant individuals, the other subpopulation will still resist invasions by tolerant mutants. Incentives are thus not enough to make a fully tolerant community.

On the other hand, Figure 10 shows that the region of $T'I|T'$ invasions does overlap with the region of $T'|II$ invasions. Thus, by providing incentives to both A and B groups, there are chances to transform a fully intolerant community into a fully tolerant one. The latter is also stable because the region where $T'T'|I$ invasions succeed does not overlap with the $T'I|T'$ region.

Figure 10 also shows that if an II community has been formed through a sequence of invasions $TT \rightarrow IT \rightarrow II$ from an original TT community, the values of ϵ and y where this happens are compatible with those for which full tolerance might be restored through (extreme, i.e., $c' = 0$) incentives. The extreme case $c' = 0$ is not very realistic, so in general one will have $0 < c' < c$, and the overlap regions are presumably smaller. But worse than that is the fact that, as soon as incentives are removed, the scenarios of Figs. 1–6 are restored and intolerance takes over again. So the effect of incentives is not permanent.

V. DISCUSSION

We have studied under which conditions intolerance can invade a tolerant community made of two distinguishable subgroups. Our results show differences between two sets of the stable and coherent strategies known in the literature of indirect reciprocity as leading eight strategies. In the first set (Groups I and II) intolerance can invade a tolerant subpopulation if the action error (which might associate to lack of resources) is high enough and/or the benefit-to-cost ratio is low. This invasion is more likely to occur if the invaded subpopulation is a minority. Once one of the two subpopulation becomes intolerant, there is no way to restore tolerance under the same economic

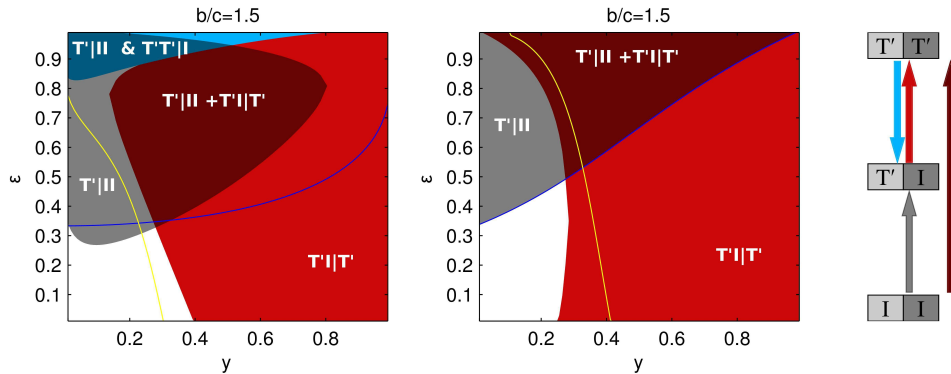


FIG. 10. (Color online.) Conditions of ϵ and y under which the invasions $T'|II$, $T'I|T'$ and $T'T'|I$ are successful for Ib (left) and IId (right) strategies. Dark (light) line shows the limit of the $I|TT$ ($IT|T$) region. Incentivated cost $c' = 0$.

conditions. Moreover, if the intolerant minority is large enough, intolerance can also invade the tolerant minority. A completely intolerant population cannot return to tolerance even if economic conditions improve.

If one subpopulation is provided incentives to help the other, tolerance can be restored for a limited region of parameters, but the other subpopulation does not turn to tolerance by invasion. To transform a fully intolerant community into a fully tolerant one both subpopulations must be stimulated to help the other. Nonetheless, as soon as these incentives are removed, intolerance can invade again. Thus, incentives have no permanent effect unless the economic conditions are simultaneously improved. Incentives may provide a temporary mean to achieve the goal, but improving the economic conditions is key to make the effect perdurable.

In the second set of strategies (Group III), a fully tolerant population is intolerance-proof. As a matter of fact, if we start off from a fully intolerant community, tolerant mutants can invade it much easier than for the first set of strategies (Groups I and II). The difference between the second set and the first set—hence the reason why they resist intolerance—is that the only way for an individual with bad reputation to achieve a good reputation is if she helps to another individual with good reputation. In other words, these strategies are less prone to forgive.

A way to interpret errors in actions is as a measure of global poverty of the society. If resources are scarce, it may happen that an individual is forced to deny help that would otherwise provide. But there may exist alternative interpretations, e.g., it can also stand as the probability that an individual is discovered when she secretly tries to cheat (defect). In this case, the results suggest that intolerance spreads easier in a society with a high proportion of known fraud. Note that in our model players do not know the reasons that may drive someone to defect, i.e., they do not know if a player defects on purpose or because she lacks the resources for it. The knowledge of these inner motivations would involve much higher cognitive processes than the simple direct observation that we are considering here.

In contrast to Nakamura and Masuda [15], we did not need to make any mean-field approximation since reputations in our model come directly from the observation of every interaction. The model we introduced is a simpli-

fied description of very complex behaviours in real world. In particular, our model considers that the whole society can be split into two classes whose differentiation is based only on one set of traits. However many different sets of traits can differentiate individuals in real life, and different people may classify individuals according to different criteria. Even the relative importance of some of these traits may change in time [38]. In complex scenarios with more classes of individuals and different ways to classify them the results will presumably be different.

The study we have conducted has further limitations. Apart from the obvious fact that using an indirect reciprocity model with binary reputation and complete information certainly oversimplifies the problem (see [39] and [32] for a discussion of these issues), the main one is that we have considered only the onset of invasion by a small fraction of mutants, but not the existence or absence of mixed equilibria (as can be done, e.g., in direct reciprocity models [40]). In other words, this study does not address the more difficult question whether there is a turn over of the population or a coexistence of strategies. And definitely simultaneous invasions by different type of mutants are excluded. There is the implicit assumption that mutants emerge at a rate slower than the time needed to reach an equilibrium. The existence of mixed equilibria thus remains an open problem.

These limitations notwithstanding, our study provides some simple clues that associate the emergence of intolerant attitudes with economic stress and the relative weight of the subgroups. Remarkably, given the simplicity of the model, economic stress is recognised as one of the main predictors for the emergence of intolerance [41]. As a matter of fact, this is a typical prediction of social ‘scapegoat’ theories [42], examples of which are the research on the increased lynching of American blacks during times of economic distress [43, 44] and the rise of fascism and anti-Semitism in Germany during the inter-war years [45, 46]. Our study also makes some predictions that are amenable to be tested in real situations, like the fact that intolerance emerges predominantly within minorities, or—perhaps more importantly—that societies with a stricter ethics are more resistant to intolerance than those indulging a relaxed attitude towards defective behavior of their members. We hope that this work stimulates empirical studies in these directions.

ACKNOWLEDGMENTS

This work is funded by the Ministerio de Economía y Competitividad (Spain) through grant PRODIEVO, and by the Comunidad de Madrid (Spain) through grant MODELICO-CM. LAM-V was supported by a postdoctoral fellowship from Alianza 4 Universidades.

Appendix A: Homogeneous populations

It will prove useful to introduce the auxiliary function

$$U(x_1, x_2) = \sum_{\alpha\beta} \chi_\alpha(x_1) \chi_\beta(x_2) P_{\alpha\beta}, \quad (\text{A1})$$

with $\chi_\alpha(x)$ as defined by (3) and

$$P_{\alpha\beta} = (1 - \epsilon) m_{\alpha\beta}(a_{\alpha\beta}) + \epsilon m_{\alpha\beta}(D) \quad (\text{A2})$$

the probability that a donor with reputation α performing the action $a_{\alpha\beta}$ on a recipient with reputation β is considered good according to the moral rule $m_{\alpha\beta}(a)$.

In a homogeneous population—where all its members belong to same type—the dynamics of the fraction of good individuals g follows the differential equation

$$\frac{dg}{dt} = U(g, g) - g. \quad (\text{A3})$$

In equilibrium $g = U(g, g)$, a quadratic equation with a unique stable solution $0 \leq g \leq 1$ [32].

For later convenience, we will consider a continuous perturbation of the above equation, namely

$$x = aU(x, x) + 1 - a, \quad (\text{A4})$$

and denote $x = p(a)$ its stable solution in the interval $[0, 1]$. Clearly $p(1) = g$, the solution of the unperturbed equation.

Another function that we will use later is the solution of the linear equation $x = aU(x, k) + 1 - a$, $0 \leq a, k \leq 1$, which can be readily obtained as

$$x = \frac{1 + a[kP_{01} + (1 - k)P_{00} - 1]}{1 - a[kP_{11} - (1 - k)P_{10} + kP_{01} + (1 - k)P_{00}]} \equiv \tilde{p}(k, a). \quad (\text{A5})$$

Appendix B: Equilibria for the different scenarios

We must obtain the equilibrium solutions of Eqs. (5)–(8) in order to compute the probabilities $\theta_{i,j}$, which depend on $x_i^{\Lambda_A \Lambda_B \Lambda_M}$. The set of Eqs. (8) is decoupled from Eqs. (5)–(7), hence $x_M^{\Lambda_A \Lambda_B \Lambda_M}$ can be calculated analytically from Eq. (8) after solving Eqs. (5)–(7). Thus one can readily obtain g_M^M and reduce Eq. (8) to a linear system of two equations in two unknowns (for example, x_M^{G*G} and x_M^{B*B}). In what follows, we describe the processes to complete our calculations for each different scenario.

1. I|TT scenario

This case considers the possible invasion of a fully tolerant resident population by a small fraction of intolerant mutants. Since residents do not distinguish classes, the resident population acts as a homogeneous population. Therefore $x_i^{\Lambda_B \Lambda_M} = x_i^{\Lambda_A \Lambda_M}$, $\theta_{B,A} = \theta_{A,B} = \theta_{A,A}$, $\theta_{B,M} = \theta_{A,M}$, and $\theta_{M,B} = 0$. The fraction of good residents is $g_B^A = x_A^A = p(1)$, as previously discussed. According to the definition of g_A^A , Eq. (4), we have

$$x_A^{G*B} = g_A^A - x_A^{G*G}, \quad (\text{B1})$$

$$x_A^{B*G} = 1 - g_A^A - x_A^{B*B}. \quad (\text{B2})$$

Then, the fraction $F_{A,B}^{\Lambda_A \Lambda_M}$ introduced in Eq. (5) can be expressed in this scenario as

$$F_{A,B}^{\Lambda_A \Lambda_M} = \sum_{\alpha\beta} \chi_\alpha(g_A^A) \chi_\beta(g_A^A) P_A^{\Lambda_A \Lambda_M}(\alpha\beta), \quad (\text{B3})$$

$$P_A^{\Lambda_A \Lambda_M}(\alpha\beta) = (1 - \epsilon) \delta(\Lambda_A, m_{\alpha\beta}(a_{\alpha\beta})) \times \delta(\Lambda_M, 1 - a_{\alpha\beta}) + \epsilon \delta(\Lambda_A, m_{\alpha\beta}(D)) \times \delta(\Lambda_M, 1). \quad (\text{B4})$$

Once we numerically solve the two Eqs. (5), Eq. (8), which corresponds to the intolerant mutants, can be solved analytically by taking into account that

$$F_{M,B}^{\Lambda_A \Lambda_M} = \sum_{\alpha\beta} \chi_\alpha(g_M^A) \chi_\beta(g_B^A) P_M^{\Lambda_A \Lambda_M}(\alpha\beta), \quad (\text{B5})$$

$$P_M^{\Lambda_A \Lambda_M}(\alpha\beta) = \delta(\Lambda_A, m_{\alpha\beta}(D)) \delta(\Lambda_M, 1). \quad (\text{B6})$$

If at equilibrium, Eq. (8) is summed over Λ_A , we find that g_M^M is given by $\tilde{p}(g_A^M, y)$. Then the set of Eqs. (8) simplifies to just two equations.

2. T|IT scenario

In this scenario a small fraction of tolerant individuals tries to invade a population where individuals of their own class are intolerant ($\theta_{A,B} = 0$), whereas those of the other class are tolerant. Since mutants and B-residents are both tolerant, they will judge every player equally. Then $x_i^{\Lambda_A \Lambda_B \Lambda_M} = x_i^{\Lambda_A \Lambda_M}$.

The dynamics of the intolerant A-strategists is described by Eq. (5), where

$$F_{A,B}^{\Lambda_A \Lambda_M} = \sum_{\alpha\beta} \chi_\alpha(g_A^M) \chi_\beta(g_B^M) P_A^{\Lambda_A \Lambda_M}(\alpha\beta), \quad (\text{B7})$$

$$P_A^{\Lambda_A \Lambda_M}(\alpha\beta) = \delta(\Lambda_M, m_{\alpha\beta}(D)) \delta(\Lambda_A, 1). \quad (\text{B8})$$

Summing Eqs. (5) over Λ_M at equilibrium we obtain $g_A^A = p(y)$. In this scenario we also need to calculate g_B^M . Since B-mutant's judgements are the same as those of B-residents, g_B^M is given at equilibrium by

$$g_B^M = (1 - y)U(g_B^M, g_B^M) + yU(g_B^M, g_A^A). \quad (\text{B9})$$

Therefore, we need to solve a system of three equations (two from Eqs. (5) plus Eq. (B9)), choosing, for instance, x_A^{G*G} , x_A^{B*B} , and g_B^M as unknowns. After solving this

system numerically, the set of Eqs. (8), where

$$F_{M,B}^{\Lambda_A \Lambda_M} = \sum_{\alpha\beta} \chi_\alpha(g_M^M) \chi_\beta(g_M^B) P_M^{\Lambda_A \Lambda_M}(\alpha\beta) \quad (\text{B10})$$

$$\begin{aligned} P_M^{\Lambda_A \Lambda_M}(\alpha\beta) &= (1 - \epsilon) \delta(\Lambda_M, m_{\alpha\beta}(a_{\alpha\beta})) \\ &\quad \times \delta(\Lambda_A, 1 - a_{\alpha\beta}) \\ &\quad + \epsilon \delta(\Lambda_M, m_{\alpha\beta}(D)) \times \delta(\Lambda_A, 1), \end{aligned} \quad (\text{B11})$$

can be solved analytically. For that we take into account that summing over Λ_A at equilibrium yields $g_M^M = \tilde{p}(y g_A^M + (1 - y) g_B^M, 1)$.

3. I|TI scenario

In this scenario a small fraction of intolerant mutants tries to invade a population where individuals of their own class are tolerant whereas those of the other class are intolerant. Hence $\theta_{B,A} = \theta_{B,M} = \theta_{M,B} = 0$.

In order to calculate g_B^A we first need to compute $x_B^{\Lambda_A \Lambda_B}$ through Eq. (6), where

$$F_{B,A}^{\Lambda_A \Lambda_B} = \sum_{\alpha\beta} \chi_\alpha(g_B^A) \chi_\beta(g_A^A) P_B^{\Lambda_A \Lambda_B}(\alpha\beta), \quad (\text{B12})$$

$$P_B^{\Lambda_A \Lambda_B}(\alpha\beta) = \delta(\Lambda_A, m_{\alpha\beta}(D)) \delta(\Lambda_B, 1), \quad (\text{B13})$$

and summing over Λ_A at equilibrium one obtains that $g_B^B = p(1 - y)$. This way, we reduce the set of Eqs. (6) to just two equations.

Now, the dynamics of the tolerant individuals is given by Eq. (5), with

$$F_{A,B}^{\Lambda_A \Lambda_M} = \sum_{\alpha\beta} \chi_\alpha(g_A^A) \chi_\beta(g_B^A) P_A^{\Lambda_A \Lambda_M}(\alpha\beta), \quad (\text{B14})$$

$$\begin{aligned} P_A^{\Lambda_A \Lambda_M}(\alpha\beta) &= (1 - \epsilon) \delta(\Lambda_A, m_{\alpha\beta}(a_{\alpha\beta})) \\ &\quad \times \delta(\Lambda_M, 1 - a_{\alpha\beta}) \\ &\quad + \epsilon \delta(\Lambda_A, m_{\alpha\beta}(D)) \times \delta(\Lambda_M, 1). \end{aligned} \quad (\text{B15})$$

The only simplification that works in this case is to take into account that $x_A^{B*G} = 1 - x_A^{G*G} - x_A^{G*B} - x_A^{B*B}$. Thus we need to solve numerically five coupled equations: three from Eqs. (5) and two from Eqs. (6), corresponding to the unknowns x_B^{G*G} , x_B^{B*B} , x_A^{G*G} , x_B^{G*B} , and x_B^{B*B} .

The dynamics of the mutants is solved from Eq. (8), where

$$F_{M,B}^{\Lambda_A \Lambda_M} = \sum_{\alpha\beta} \chi_\alpha(g_M^A) \chi_\beta(g_B^A) P_M^{\Lambda_A \Lambda_M}(\alpha\beta), \quad (\text{B16})$$

$$P_M^{\Lambda_A \Lambda_M}(\alpha\beta) = \delta(\Lambda_A, m_{\alpha\beta}(D)) \delta(\Lambda_M, 1). \quad (\text{B17})$$

Summing over Λ_A at equilibrium yields $g_M^M = \tilde{p}(g_A^M, y)$, and transforms the last equation to just two linear equations.

4. T|II scenario

In this last scenario a small fraction of tolerant mutants tries to invade a completely intolerant population. Then $\theta_{A,B} = \theta_{B,A} = \theta_{B,M} = 0$.

The dynamics for the intolerant A-residents is described by Eq. (5) with

$$F_{A,B}^{\Lambda_A \Lambda_M} = \sum_{\alpha\beta} \chi_\alpha(g_A^M) \chi_\beta(g_B^M) P_A^{\Lambda_A \Lambda_M}(\alpha\beta), \quad (\text{B18})$$

$$P_A^{\Lambda_A \Lambda_M}(\alpha\beta) = \delta(\Lambda_M, m_{\alpha\beta}(D)) \delta(\Lambda_A, 1). \quad (\text{B19})$$

Summing over Λ_M at equilibrium yields $g_A^A = p(y)$.

In order to calculate g_B^M , we need first to compute $x_B^{\Lambda_B \Lambda_M}$. The dynamics for these fractions is described by Eq. (7), where

$$F_{B,A}^{\Lambda_B \Lambda_M} = \sum_{\alpha\beta} \chi_\alpha(g_B^M) \chi_\beta(g_A^M) P_B^{\Lambda_B \Lambda_M}(\alpha\beta), \quad (\text{B20})$$

$$P_B^{\Lambda_B \Lambda_M}(\alpha\beta) = \delta(\Lambda_M, m_{\alpha\beta}(D)) \delta(\Lambda_B, 1), \quad (\text{B21})$$

and summing again over Λ_M we obtain that $g_B^B = p(1 - y)$. Thus we have four coupled equations (two from Eqs. (6) and two from Eq. (7)) for the unknowns x_A^{G*G} , x_A^{B*B} , x_B^{G*G} , and x_B^{B*B} , which need to be solved numerically.

On the other hand, the dynamics of the tolerant mutants is given by Eq. (7), where

$$F_{M,B}^{\Lambda_A \Lambda_M} = \sum_{\alpha\beta} \chi_\alpha(g_M^M) \chi_\beta(g_B^M) P_M^{\Lambda_A \Lambda_M}(\alpha\beta), \quad (\text{B22})$$

$$\begin{aligned} P_M^{\Lambda_A \Lambda_M}(\alpha\beta) &= (1 - \epsilon) \delta(\Lambda_M, m_{\alpha\beta}(a_{\alpha\beta})) \\ &\quad \times \delta(\Lambda_A, 1 - a_{\alpha\beta}) \\ &\quad + \epsilon \delta(\Lambda_M, m_{\alpha\beta}(D)) \delta(\Lambda_A, 1). \end{aligned} \quad (\text{B23})$$

This set of equations can be reduced to just two because summing over Λ_A at equilibrium yields $g_M^M = \tilde{p}(y g_A^M + (1 - y) g_B^M, 1)$.

[1] W. D. Hamilton, J. Theor. Biol. **7**, 17 (1964).
[2] W. G. Sumner, Boston. Ginn (1906).
[3] R. A. LeVine and D. T. Campbell, New York: John Wiley (1972).
[4] L. A. Hirschfeld, Cambridge, MA: MIT Press (1996).
[5] H. Bernhard, U. Fischbacher, and E. Fehr, Nature **442**, 912 (2006).
[6] D. Haig, Proc Natl Acad Sci USA **93**, 6547 (1996).
[7] L. Keller and K. G. Ross, Nature **394**, 573 (1998).

[8] D. C. Queller, E. Ponte, S. Bozzaro, and J. E. Strassmann, Science **299**, 105 (2003).
[9] K. Summers and B. Crespi, Proc R Soc Lond B **272**, 643 (2005).
[10] B. Sinervo, A. Chaine, J. Clobert, R. Calsbeek, L. Hazard, L. Lancaster, A. G. McAdam, S. Alonzo, G. Corrigan, and M. E. Hochberg, Proc Natl Acad Sci USA **103**, 7372 (2006).
[11] J. J. Ray and F. H. Lovejoy, Journal of Social Psychology

- 126**, 563 (1986).
- [12] N. Struch and S. H. Schwartz, *Journal of Personality and Social Psychology* **56**, 346 (1989).
 - [13] E. Cashdan, *Current Anthropology* **42**, 760 (2001).
 - [14] N. Masuda, *J. Theor. Biol.* **311**, 8 (2012).
 - [15] M. Nakamura and N. Masuda, *BMC Evol. Biol.* **12**, 213 (2012).
 - [16] K. Oishi, T. Shimad, and N. Ito, arXiv:1211.7164 (2012).
 - [17] R. Sugden, *The economics of rights, co-operation and welfare* (Basil Blackwell, Oxford, 1986).
 - [18] R. D. Alexander, *The Biology of Moral Systems* (New York: Aldine de Gruyter, 1987).
 - [19] R. L. Trivers, *Q. Rev. Biol.* **46**, 35 (1971).
 - [20] M. Dufwenberg, U. Gneezy, W. Güth, and E. van Demme, *Homo Oeconomicus* **18**, 19 (2001).
 - [21] M. Milinski, D. Semmann, and H.-J. Krambeck, *Nature* **415**, 424 (2002).
 - [22] K. Panchanathan and R. Boyd, *Nature* **432**, 499 (2004).
 - [23] D. Semmann, H.-J. Krambeck, and M. Milinski, *J. Behav. Ecol. Sociobiol.* **56**, 248 (2004).
 - [24] S. Suzuki and E. Akiyama, *J. Theor. Biol.* **245**, 539 (2007).
 - [25] R. Bshary and A. S. Gutter, *Nature* **441**, 975 (2006).
 - [26] G. E. Bolton, E. Katok, and A. Ockenfels, *Manage. Sci.* **50**, 1587 (2005).
 - [27] C. Keser, *IBM Syst. J.* **43**, 498 (2002).
 - [28] M. A. Nowak and K. Sigmund, *Nature* **393**, 573 (1998).
 - [29] H. Ohtsuki and Y. Iwasa, *J. Theor. Biol.* **231**, 107 (2004).
 - [30] H. Brandt and K. Sigmund, *J. Theor. Biol.* **231**, 475 (2004).
 - [31] M. A. Nowak and K. Sigmund, *Nature* **437**, 1291 (2005).
 - [32] L. A. Martinez-Vaquero and J. A. Cuesta, *Phys. Rev. E* **87**, 052810 (2013).
 - [33] H. Ohtsuki and Y. Iwasa, *J. Theor. Biol.* **239**, 435 (2006).
 - [34] O. Leimar and P. Hammerstein, *Proc. R. Soc. Lond. B* **268**, 745 (2001).
 - [35] K. Panchanathan and R. Boyd, *J. Theor. Biol.* **224**, 115 (2003).
 - [36] M. A. Fishman, *J. Theor. Biol.* **225**, 285 (2003).
 - [37] A. Lotem, M. A. Fishman, and L. Stone, *Nature* **400**, 226 (1999).
 - [38] R. Kurzban, J. Tooby, and L. Cosmides, *Proc. Natl. Acad. Sci. USA* **98**, 15387 (2001).
 - [39] S. Uchida and K. Sigmund, *J. Theor. Biol.* **263**, 13 (2010).
 - [40] L. A. Martinez-Vaquero, J. A. Cuesta, and A. Sánchez, *PLoS ONE* **7**, e35135 (2012).
 - [41] J. L. Gibson, *Brit. J. Pol. Sci.* **32**, 309 (2002).
 - [42] H. Tajfel, *Human Groups and Social Categories* (Cambridge University Press, Cambridge, 1981).
 - [43] C. I. Hovland and R. S. Sears, *J. Psych.* **9**, 301 (1940).
 - [44] J. Dollard, N. E. Miller, L. W. Doob, O. H. Mowrer, and R. R. Sears, *Frustration and Aggression* (Yale University Press, New Haven, Connecticut, 1939).
 - [45] J. Levin and W. C. Levin, *The Functions of Discrimination and Prejudice* (Harper and Row, New York, 1982).
 - [46] M. Billig, *Fascists: A Social Psychological View of the National Front* (Academic Press, London, 1978).